What have the Romans (and Germans) ever done for us? or There are real applications for Description Logics some of which even take advantage of their reasoning services

Carole Goble
Information Management Group
Department of Computer Science
University of Manchester
email:carole@cs.man.ac.uk
http://img.cs.man.ac.uk/carole

One of the nagging worries of the Description Logic community is what applications are really there that make use of their efforts. So the hunt is on for some sort of "killer application" that can demonstrate an improvement, or success where previously there was none, by the use of Description Logic based reasoning.

As an application builder outside the Description Logic community, who tries to deploy Description Logics as part of a portfolio of techniques in real applications, I hope to reassure this community that, yes, there are useful applications of their endeavours - that it is all worthwhile. These applications may not necessarily guarantee the silver bullet that the Description Logic community yearns for, and maybe they don't all use the "full power" of a Description Logic, but the message is: the Description Logic helped and was thus helpful, and we believe that it could potentially help more.

The work I will discuss, using practical illustrative examples scattered throughout the talk, chiefly concerns applications in:

- Bioinformatics, specifically intelligent retrieval and multi-source integration through the TAMBIS project [1, 2] and improved database annotation in the Irbane and Pre-PRINTS projects;

- Medical Informatics, specifically medical controlled vocabularies through the GALEN project [8, 9] and intelligent clinical data entry experienced through the PEN&PAD projects [7];

- Digital Libraries, specifically picture collection management, classification based retrieval, metadata management, and thesaurus construction, experienced through the STARCH project [3].

I will also refer to the promise Description Logics hold for realising "The Semantic Web", currently being explored in the OIL activity, and the subject of a paper in this workshop [6].

These applications all make extensive use of DL models of their domains of interest. The models we have built at Manchester have been large, collaborative and incremental - the OpenGALEN model of basic anatomy alone contains 5000 concepts, the model of surgical procedures some 15,000. Comprehensive clinical terminologies require at least 250,000 concepts. The TAMBIS bio-ontology is smaller at some 1500 concepts. The use of DLs and DL based tools has been crucial to the management of these large endeavours.

Both Bioinformatics and Medical Informatics have become increasingly important with the escalating amount of data generated by the scientific community that must be shared, integrated and understood. Both disciplines fundamentally rely on finding and linking evidence from different information sources, experiments, episodes, and workers; and then analysing and comparing this evidence, recognising the presence or absence of relationships, properties and patterns in general. Both therefore rely on the capture, retrieval and analysis of high quality metadata and knowledge.

Andy Brass, one of our closest bioinformatics collaborators, is fond of saying that Bioinformatics is a knowledge-based discipline. Many predictions, and interpretations, of data in biology are made by comparing the data in hand against existing knowledge. For example, consider the problem of predicting protein function from sequence. If an unknown protein sequence A resembles the well-characterised protein sequence B, and we know the function of B then perhaps A's function can be inferred from the type of similarities found. The key difference therefore between "knowledge-based" and "axiomatic" disciplines is the role played by the knowledge base of past experience. The challenge and the skill in biology is to make use of this knowledge in the most effective way. Thus a rich and extensible model of the concepts, properties, relationships and constraints that can be used to infer new implied relationships, or identify inconsistent definitions, is considered to be an extremely valuable resource.

Exploiting reasoning to discover knowledge and drive further scientific investigation is potentially a very exciting application for DLs. One biological discovery directed by an implied relationship in a knowledge model doesn't just produce a tidier ER schema for a telecoms database, or an appropriate configuration for a computer network - it may be a step on the way to discovering a new Viagra or Prozac.

I will conclude with some remarks about our experiences of the difficulties in applying the DL approach, our solutions (if any) and what the applications community needs from the DL community. These remarks include:

- Explanation: why was a concept unsatisfiable? Nothing is more frustrating or discouraging than a negative result without a hint of why, especially when your model is large, complicated and built by a bunch of other people.

- The expressivity that we have found essential and non-essential. Transitive roles, inverse roles and role hierarchies are heavily used, as are general inclusion axioms. Part-whole reasoning is ubiquitous in medical terminologies (all that anatomy) and to some extent in structural biology. Values or (simple) concrete domains are important but perhaps surprisingly, we have happily done without number restrictions. More interesting perhaps is that we have done without A-box reasoning - everything we have done requires T-box reasoning alone, although we do have concepts that could perhaps be classed as pseudo-individuals. In TAMBIS our A-box is effectively the bioinformatic data resources we are querying over; in STARCH the only individual is the picture descriptor itself, isolated from any other individual.

- The addition of frame-like behaviour to drive our data entry and retrieval interfaces - both depend upon being able to determine which roles are reasonable to apply to a particular concept. We call this sanctioning, and a paper on how it was implemented in FaCT is in this workshop [5].

- The need for tools that can cope with large models. DL systems that have clean APIs (such as the CORBA-FaCT interface [4]), and can form components of larger systems with additional services such as sanctioning, lexicons, extrinsics etc. Tools for building complex and highly connected ontologies collaboratively and over long time frames. Experiences in GALEN have shown the value of authoring tools for domain experts of a simplified "intermediate representation", which is then translated into the Description Logic, which is then relegated to the status of an "assembly language".

Some final comments. Firstly, an application developer toying with the idea of using a Description Logic in their application is often told that "if you aren't using the full power of the reasoning services, then you aren't using it properly, so it would be better if you didn't use it at all". I know I have. If the developer is planning to use a hammer where a spanner is the appropriate tool, then fair enough. However, if the DL would have done as well as anything else, then this is like saying "because you only use a relational database to hold a couple of tables and you aren't using stored procedures or correlated sub-queries, you aren't using it properly so don't use it at all". This seems like shooting yourselves in the foot. Embrace your users, encourage them to come and play, and they might even buy into using your reasoning.

Secondly, take advantage of the characteristics of real applications. FaCT (which we adopted in TAMBIS, STARCH and OIL), and GRAIL (a primitive DL with little expressivity used in the earlier TAMBIS, PEN&PAD and GALEN) both take advantage of the fact that in the real world, most of the scary monsters lurking in the intractability bushes either don't actually manifest themselves or can be worked around.

Finally, the power of DLs is really only realised when the reasoning services have something to bite on - rich models, metadata descriptions and queries that make use of term constructors, defined concepts and axioms. In contrast, domain modellers tend to think in terms of asserted primitive taxonomies, or frame-like structures. A major challenge is to build modelling tools that ensure the Description in Description Logics.

Acknowledgements

I'd like to acknowledge Alan Rector for his insights and experiences in developing medical terminologies in Description Logics that he willingly shared with me. I'd also like to thank Pat Baker and Robert Stevens, our bio-ontologists, for their patience and examples.

References

[1] Baker P.G., Brass A., Bechhofer S., Goble C.A., Paton N.W., Stevens R. *TAMBIS — Transparent Access to Multiple Bioinformatics Information Sources. An Overview.* In Proceedings of Information Systems in Molecular Biology, June 1998.

[2] Baker P.G., Goble C.A., Bechhofer S., Paton N.W., Stevens R. and Brass A. *An Ontology for Bioinformatics Applications.* In Bioinformatics Vol 15, No 6, pp: 510-520, (1999).

[3] Bechhofer S. and Goble C.A. *Classification Based Navigation and Retrieval for Picture Archives.* In Database Semantics, Semantic Issues in Multimedia Systems. Proceedings of IFIP TC2/WG2.6 Eighth Working Conference on Database Semantics (DS-8), Rotorua, New Zealand, January 4-8, 1999 Eds: Robert Meersman, Zahir Tari, Scott Stevens Kluwer Academic Publishers, pp 291–310.

[4] Bechhofer S., Horrocks I., Patel-Schneider P.F., Tessaris S.. *A Proposal for a Description Logic Interface.* In DL99, International Workshop on Description Logics, Linköping, Sweden, 1999.

[5] Bechhofer S, Horrocks I., *Driving User Interfaces from FaCT*, In DL2000, International Workshop on Description Logics, Aachen, Germany, 2000.

[6] Decker S, Fensel D, van Harmelen F, Horrocks I, Melnik S, Klein M, Broekstra J, *Knowledge Representation on the Web.* In DL2000, International Workshop on Description Logics, Aachen, Germany, 2000.

[7] Nowlan WA, Rector AL, Kay S, Goble CA, Horan B, Howkins TJ and Wilson. *A PEN & PAD: A Doctor's Workstation with Intelligent Data Entry and Summaries*. In RA Miller (ed) Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care, SCAMC 90, IEEE Computer Society Press, Los Alamitos, California, 1990, pp 941-942 (1990).

[8] Rector A.L. Compositional Models of Medical Concept: Towards Re-Usable Application - Independent Medical Terminologies (PS) Knowledge and Decisions in Health Telematics, (Eds) P Barahona and J P Christensen, IOS Press 1994. pp 109 - 114

[9] Rector A.L., Bechhofer S., Goble C.A., Horrocks I., Nowlan W.A., and Solomon W.D. *The GRAIL Concept Modelling Language for Medical Terminology.* Journal of Artificial Intelligence in Medicine, Kluwer Publishing, Vol 9, pp: 139-171, (1997).